Mejora del análisis con Boxplot

D

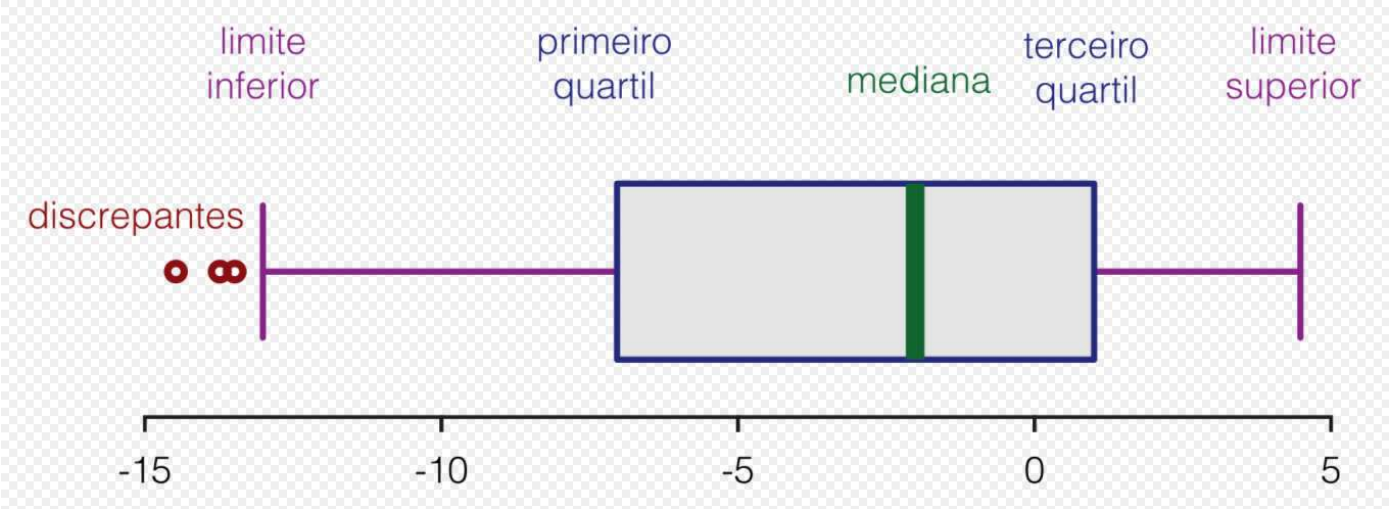
danpsiqueira

24 de Marzo

¿Qué es un BoxPlot?

Un BoxPlot (o diagrama de caja, en traducción libre) muestra la distribución cuantitativa de los datos de una manera que facilita la comparación entre las variables, o a través de los niveles categóricos de las variables.

Esta caja ("box") muestra los cuartiles del conjunto de datos mientras que los "whiskers" muestran el resto de la distribución, excepto los puntos que se denominan valores atípicos.

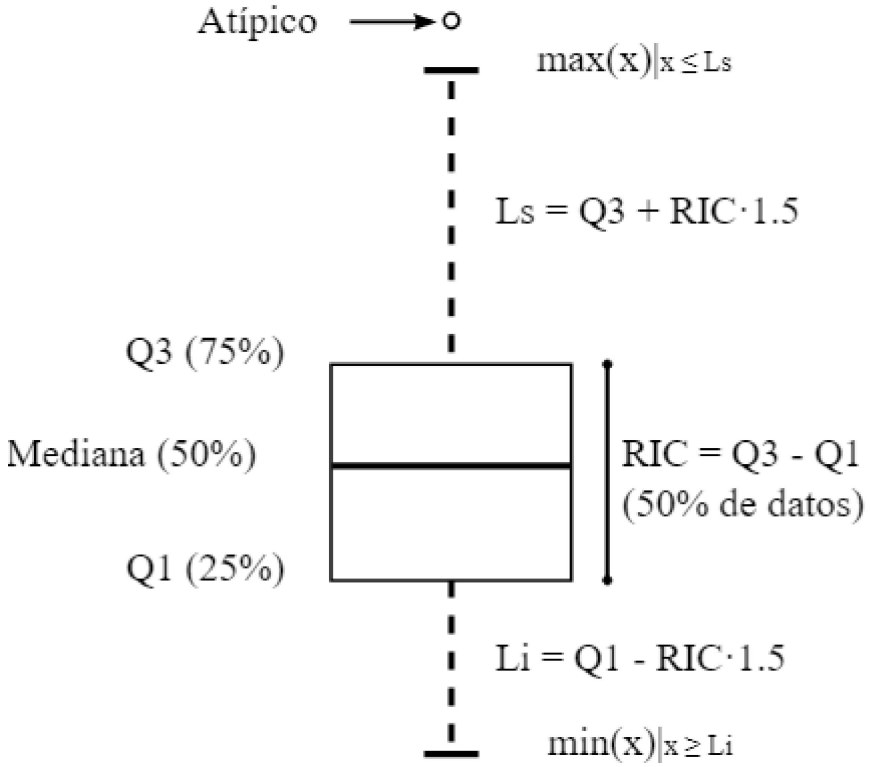


La línea en el centro de la figura en verde, en nuestro caso, representa la mediana. La línea azul que está en el borde izquierdo de la figura representa el 25 % de mi distribución, y la línea que está en el borde derecho representa el 75 % de la distribución de datos.

En las líneas moradas que están en los extremos de ambos lados, dejan la figura central y muestran una delimitación, tenemos el inicio y el final del área de los cuartiles. Ahora, observe que hay algunos puntos después de este límite en el lado izquierdo. Estos puntos se denominan valores atípicos, es decir, son puntos “aislados” en nuestra distribución.

Para saber hasta dónde llegan los whiskers, que son esos “bigotes de gato” que delimitan los cuartiles y marcan dónde empiezan los outliers, hagamos un cálculo. Una vez que encontramos la mediana, el punto central de la distribución, sabemos que tenemos el 50% de la distribución a la izquierda y el 50% a la derecha. Después de eso, dividamos los lados por la mitad nuevamente, para obtener dos partes de 25% a la derecha y 25% a la izquierda.

Ahora que tenemos los 4 cuartiles, cada uno con un 25%, podemos dibujar la caja, o “box”, y delimitar con el 2° y 3° cuartiles. Entonces, para encontrar la delimitación de los wiskers, multipliquemos la distancia desde el segundo cuartil hasta la mediana por 1,5. Hacemos el mismo procedimiento con la distancia del 3er cuartil a la mediana: multipliquemos por 1,5.



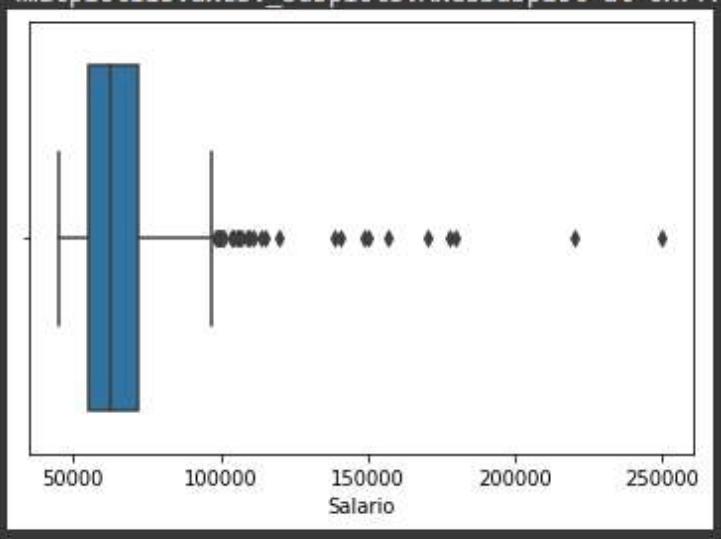
En resumen, el diagrama de caja nos ayuda a visualizar la distribución de los datos divididos en cuartiles. Además, muestra dónde están más concentrados los datos y si hay valores atípicos fuera de nuestros cuartiles.

¿Cómo generar un BoxPlot usando Python?

Para generar el Boxplot en Python usaremos la librería Seaborn. También podríamos generar el Boxplot a través de otros métodos, pero una de las ventajas de usar Seaborn es que será más bonito, claro y presentable.

```
[12] sns.boxplot(x = dados['Salario'])
```

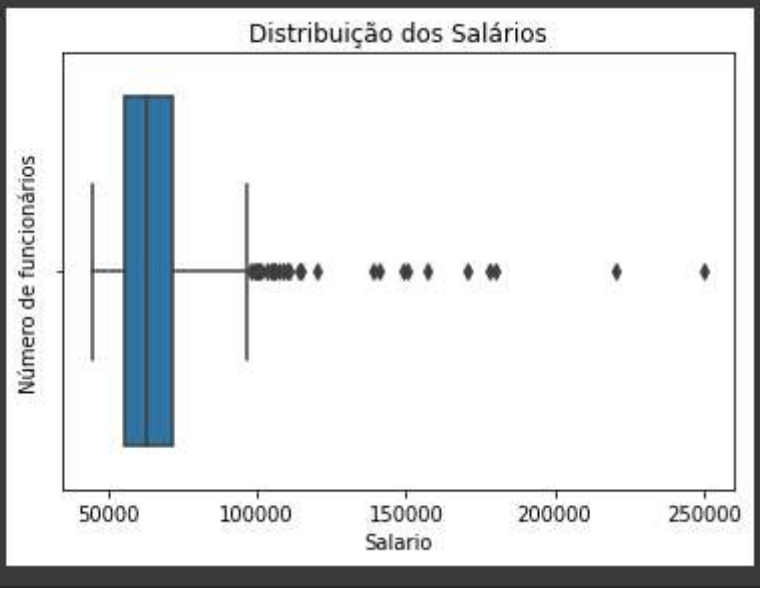
```
<matplotlib.axes._subplots.AxesSubplot at 0x7f70d22902d0>
```



Para cambiar los ejes x e y de nuestro diagrama podemos usar la librería matplotlib, a través del código: `import matplotlib.pyplot as plt sns.boxplot(x = data['Salary']) plt.ylabel('Number of employee') plt.show()`

```
import matplotlib.pyplot as plt
sns.boxplot(x = dados['Salario'])

plt.title('Distribuição dos Salários')
plt.ylabel('Número de funcionários')
plt.show()
```



Realización de análisis e hipótesis con BoxPlot

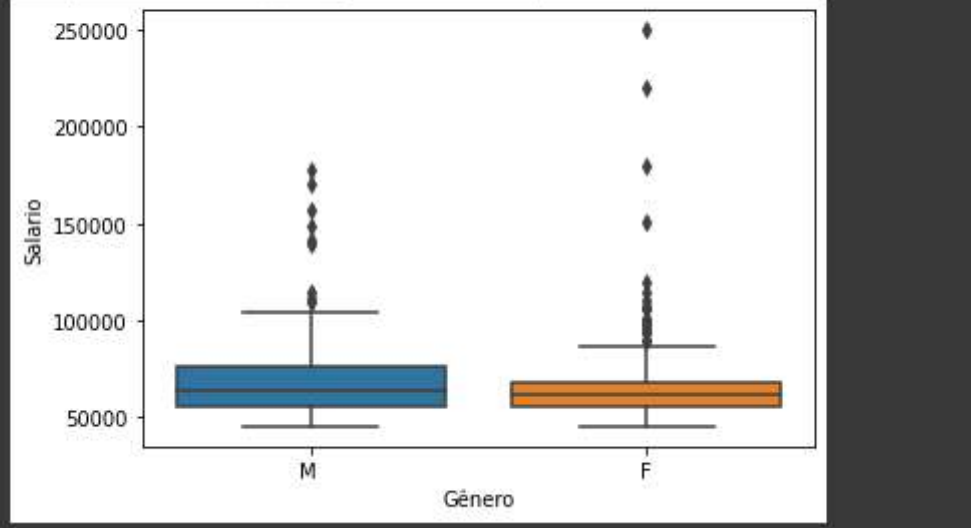
Analizando este boxplot, notamos que la mayoría de los datos están un poco por encima de 50000 y antes de 100000. Si analizamos el histograma generado con estos mismos datos, nos damos cuenta de que la información complementa el uno al otro. Tenemos pocos datos (outliers) que estén por encima de 100000 por año.

Podemos ir más allá en el uso de diagramas de caja. Podemos usar más de un diagrama de caja con diferentes categorías (columnas) para hacer análisis de comparación entre ellos.

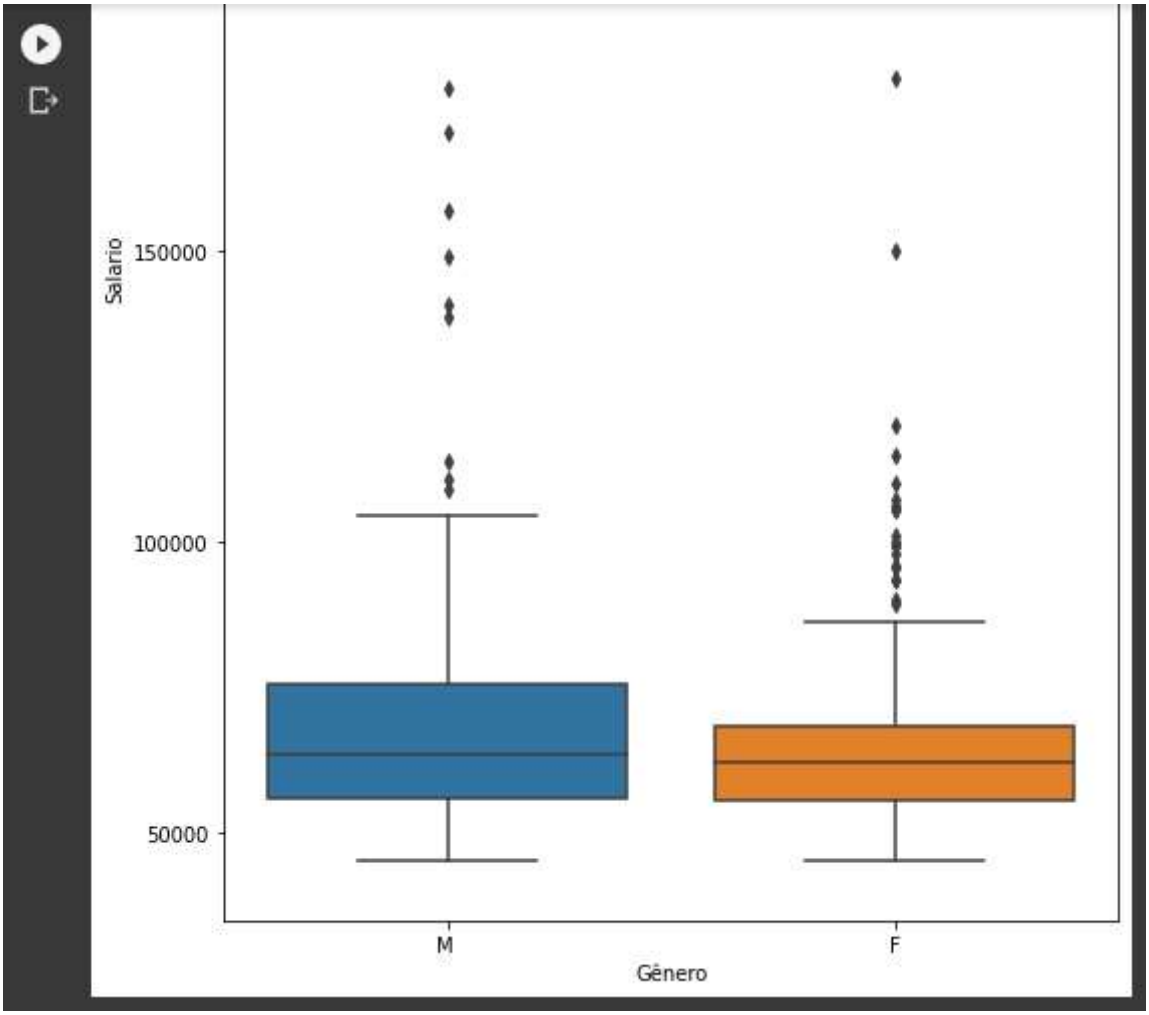
Hagamos un experimento: analicemos a qué tipo de conclusiones podemos llegar comparando el salario anual con el género del empleado. Para hacer esto, generemos el boxplot con el siguiente código:

```
sns.boxplot(data=dados, x='Gênero', y='Salario')
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7faaadd4f590>
```



Como en la categoría de género femenino tenemos más valores atípicos, el gráfico se aplanó un poco. Usemos matplotlib para mejorar un poco la visualización:



Analicemos un poco estos diagramas de caja (boxplots). No pudimos responder directamente si los hombres ganan más que las mujeres, al menos solo con este paso, ya que notamos que la mediana de las dos gráficas de caja está muy cerca una de la otra. Pero pudimos analizar que el 3er cuartil del género femenino termina mucho antes que el género masculino, lo que nos puede decir que a medida que aumenta el salario anual, tenemos una mayor concentración de personas del género masculino. También notamos que hay muchos más valores atípicos, datos atípicos en la distribución, de mujeres que tienen salarios altos.

¡Ahora es tu turno! Siguiendo estos mismos pasos, analice un conjunto de datos. Puede analizar, por ejemplo, los salarios de alguna empresa. Usa el diagrama de caja y cuéntanos en los comentarios qué conclusiones pudiste sacar.

Daniel Siqueira

Daniel es instructor en Data School y enseña Matemáticas, Física, Química e Inglés. Tiene una verdadera pasión por aprender cosas y temas nuevos, y transmitir sus conocimientos.

Cursos de Data Science

ARTÍCULOS DE TECNOLOGÍA > DATA SCIENCE

- Estudia las 24 horas, los 7 días de la semana
- Foro y comunidad exclusiva para resolver tus dudas
- Acceso a todo el contenido de la plataforma por 6 meses

¡QUIERO EMPEZAR A ESTUDIAR!

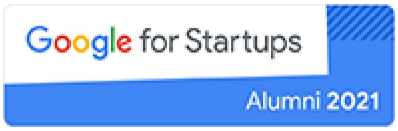
Paga en moneda local en los siguientes países

ANUAL

US$79,90

un solo pago de US$79,90

- 218 cursos
- Videos y actividades 100% en Español
- Certificado de participación
- Estudia las 24 horas, los 7 días de la semana
-



Fuimos unas de las 7 startups seleccionadas por Google For Startups en participar del programa Growth Academy en 2021

POWERED BY

CURSOS

Cursos de Programación
Lógica de Programación | Java

Cursos de Front End
HTML y CSS | JavaScript | React

Cursos de Data Science
Data Science | Machine Learning | Excel | Base de Datos | Data Visualization | Estadística

Cursos de DevOps
Docker | Linux

Cursos de Innovación y Gestión
Productividad y Calidad de Vida | Transformación Ágil | Marketing Analytics | Liderazgo y Gestión de Equipos | Startups y Emprendimiento